Research Article

Statistics in Medicine

Received 25 March 2011, Accepted 11 May 2011 Published online 22 July 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4305

Monitoring binary outcomes using risk-adjusted charts: a comparative study

Edit Gombay,[a] Abdulkadir A. Hussein[b*†] and Stefan H. Steiner[c]

Monitoring binary outcomes when evaluating health care performance has recently become common. Classical statistical methodologies such as cumulative sum (CUSUM) charts have been refined and used for this purpose. For instance, the risk-adjusted CUSUM chart (RA-CUSUM) for monitoring binary outcomes was proposed for monitoring 30-day mortality following cardiac surgery. The RA-CUSUM inherits optimality properties of the original CUSUM charts in the sense of signaling early when there is change. However, although the RA-CUSUM is a powerful monitoring tool, it will always eventually signal a change with probability 1 even when there is no real change. In other words, the probability of a type I error for the RA-CUSUM is 1. It also turns out that, because of the skewed distribution of the run lengths of the RA-CUSUM, the median is often well below the mean, and as a consequence more than half of all its false alarms occur before the designed average run length. In addition, when the change to be detected occurs at a later time in the series of observations being monitored, the rate of false alarms increases, and the RA-CUSUM may not be appropriate. Therefore, if the price of false alarms is high, it is preferable to use methods that control the rate of false alarms. In this paper, we propose alternative sequential curtailed and risk-adjusted charts that control the type I error rate in the context of monitoring 30-day mortality following cardiac surgery. We explore the merits of each of these methodologies in terms of average run lengths as well as in terms of type I error probabilities, and we compare them to the RA-CUSUM chart. We illustrate the methodologies by using data on monitoring performance of seven surgeons from a medical center.

Keywords: CUSUM; risk-adjusted; curtailed sequential tests; change point

1. Introduction

Recently, monitoring health care outcomes such as post-operative mortality rates has become common, and new statistical methodologies such as risk-adjusted cumulative sum control charts (RA-CUSUM) charts have emerged for this purpose. Risk-adjusted charts were used for instance in monitoring the risk of Down's Syndrome among Norwegian newborn babies of mothers over 30 years of age [1]; in the report of Bristol Royal Infirmary Inquiry [2], which monitored the annual mortality rates for open-heart surgery on children under 1 year of age; in the case of the general practitioner Dr. Harold Shipman, who was convicted for murdering more than 200 of his patients [3]; and in monitoring incidence of congenital malformations after the Thalidomide Tragedy of the 1960s.

When monitoring in the medical context, it is often necessary to account for variations in patient conditions which lead to differential prior risks at the time of treatment. That is, while monitoring the adverse outcomes of medical procedures, we must take into account the heterogeneity of the baseline risk in order to avoid unwanted false-alarms, respond quickly to the changes, and release the clinicians from unjustified accusation as a result of treating high-risk patients. Hence, the notion of risk adjustment was introduced into monitoring, which means adjustment for the type of procedure and patient conditions (risk factors) collectively referred to as case-mix.

[a] *Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada*
[b] *Department of Mathematics and Statistics, University of Windsor, Windsor, ON, Canada*
[c] *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada*
**Correspondence to: Abdulkadir A. Hussein, Department of Mathematics and Statistics, University of Windsor, 401 Sunset Ave., Windsor, ON, N9B 3P4, Canada.*
[†] *E-mail: ahussein@uwindsor.ca*

Cumulative monitoring over time is needed for rare adverse events rather than annual routine examination of the data [4]. Therefore, statistical process control methodologies, in particular RA-CUSUM, have been recommended for monitoring performance in health care. In addition to RA-CUSUMs, there are many other methods that used risk-adjusted monitoring such as (i) variable life-adjusted display of [5, 6], (ii) cumulative risk-adjusted mortality [7], (iii) sequential probability ratio test (SPRT) of [4], (iv) resetting SPRT of [4], (vi) Shewhart p-charts of [8], and (vii) the sets method of [9]. For further examples and discussions of the above charts, the reader is referred to Woolall [10], Grigg and Spiegelhalter [11, 12], and Lai [13].

The most commonly used risk-adjusted chart is the RA-CUSUM. Because of the open-ended (indefinite) monitoring in RA-CUSUM, the probability of signaling alarm when there is no real problem (type I error) is 1. Also, the distribution of time to alarm (known also as run length) is exponential-type skewed distribution, and therefore, although the average run length (ARL) may be long when there is no problem, the likelihood of stopping much earlier than the calculated theoretical ARL is high. Furthermore, these classical CUSUM procedures have recently come under serious theoretical scrutiny. Mei [14] has shown that the control limits of the monitoring algorithm may work quite differently from expected under many scenarios, resulting in performances that are not well understood. To overcome this problem, one may employ curtailed sequential monitoring procedures based on score statistics as in [15]. In this paper, we propose four sequential curtailed and risk-adjusted charts by using score statistics. We perform Monte Carlo simulations to explore the merits of each of these methods in terms of ARLs as well as in terms of type I probabilities. We also compare the proposed methods to the RA-CUSUM chart. We illustrate the methodologies by using data on monitoring performance of seven surgeons from a cardiac surgery center in the UK.

In Section 2, we describe the proposed sequential curtailed and risk-adjusted procedures. In Section 3, we use baseline parameters taken from the data set on cardiac surgery outcomes reported in [16], and we set up Monte Carlo simulations to assess the merits of each of the methods in terms of ARL and type I errors. In Section 3.2, we apply the methods to the cardiac surgery data, and in Section 4, we provide some discussion and recommendations.

2. The risk-adjusted charts

2.1. *Concept of risk-adjusted*

One of the first risk-adjusted charts was the RA-CUSUM chart proposed for Down's syndrome assessment in [1] and later for monitoring surgical outcomes in [16].

Risk adjustment accounts for patients' prior risk factors when evaluating treatment risk. In the context of mortality after surgical operation, let π_t be the probability of an adverse event for the tth patient. This probability is a function of the patient's covariates. In cardiac surgery, we use a logistic model based on Parsonnet score (see [17]) so that

$$\log\left(\frac{\pi_t}{1-\pi_t}\right) = \alpha + \beta x_t, \tag{1}$$

where x_t is the Parsonnet score of the patient. We estimate parameters α and β from historical data sets.

We monitor surgical outcomes by testing the null hypothesis H_0 against the alternative H_A, which essentially represent the in-control and out-of-control situations, respectively. We base these hypotheses on odds ratios because each patient has a different baseline risk level. For a given estimated risk of failure equal to π_t, the odds of failure is $\frac{\pi_t}{1-\pi_t}$.

Risk-adjusted charts use varying π_t in the patient population in sequentially testing the hypotheses,

$$H_0 : R_t = \frac{\pi_t^0/(1-\pi_t^0)}{\pi_t/(1-\pi_t)} = R_0$$

versus

$$H_A : R_t = \frac{\pi_t^1/(1-\pi_t^1)}{\pi_t/(1-\pi_t)} = R_A, \tag{2}$$

where π_t^0 is the probability of an adverse outcome for an in-control process, and R_0 is the odds ratio of the odds of an adverse outcome for an in-control process to the odds of an adverse outcome after risk

adjustment for patient t. Similarly, π_t^1 is the probability of an adverse outcome for an out-of-control process with R_A being the odds ratio of the odds of an adverse outcome for an out-of-control process to the odds of an adverse outcome after risk adjustment for patient t. Usually, $R_A > R_0$ which indicates an increase in the failure rate. If the estimated risk π_t is based on the current conditions as in [16], then we may set $R_0 = 1$.

We can formulate the above hypotheses in terms of a more general change point detection terminology by setting

$$H_0: \ R_t = R_0, \ t = 1, 2, \ldots, n$$

$$H_A: \ R_t = R_0, \ t = 1, 2, \ldots, \tau, \ \ R_t = R_A, \ \ t = \tau + 1, \tau + 2, \ldots, n, \tag{3}$$

where τ is the change point.

The patient after whom the change started (the τth patient) is often unknown. If the sample size is indefinite (that is, $n = \infty$) and the change under the alternative is assumed to have begun with the first operated patient (that is, $\tau = 1$), then we have the hypotheses given in Equation (3) previously, and methodologies used for testing such hypotheses are known as sequential tests. The RA-CUSUM chart proposed in [16] is an example of this type of method.

We can generalize the hypotheses in Equation (3) by not specifying the alternative value R_A. This allows detection of any change regardless of its magnitude. Thus, Equation (3) becomes

$$H_0: \ R_t = R_0, \ t = 1, 2, \ldots, n$$

$$H_A: \ R_t = R_0, \ t = 1, 2, \ldots, \tau, \ \ R_t \neq R_0, \ \ t = \tau + 1, \tau + 2, \ldots, n. \tag{4}$$

Such alternative hypotheses with non-specific alternatives are also known as composite alternative hypotheses. Lai [13] discussed the use of generalized likelihood ratio (GLR) in change point detection procedures for this type of composite hypotheses. However, we find that standardized score statistics are more interpretable than the GLR procedures. Therefore, in the next section we provide change-detection procedures based on score statistics in the context of surgeon performance monitoring.

2.2. *Risk-adjusted and truncated sequential methods*

First, we describe four truncated and risk-adjusted sequential tests as well as the RA-CUSUM method for monitoring changes in surgical performances. Sequential tests and sequential change-detection methods are closely related. In sequential tests, we assume the change to be at the first observation, so it is a special case of change-detection methods where the change can be at any point in time. In practice, we will truncate all sequential tests at some point, and in many applications, as in clinical trials for example, truncation is desired. In fact, soon after the first open-ended classical sequential tests of [18] were defined, efforts began to produce their truncated (curtailed) versions, where testing continued until a maximum sample size n is reached.

Consider the null and alternative hypotheses in Equation (4) and define the standardized score statistic process at epoch t as

$$S_t = \sum_{i=1}^{t} (y_i - \pi_i) \, / \sqrt{\pi_i (1 - \pi_i)}, \tag{5}$$

where $y_i = 1$ if an adverse event occurred to the ith patient and 0 otherwise, the probability of an adverse event being $\pi_i = \exp(\alpha + \beta x_i)/(1 + \exp(\alpha + \beta x_i))$, where x_i is the Parsonnet score of the ith patient and $t = 1, \ldots, n$. Terms in the above sum are the standardized difference of observed and expected values of the adverse event indicator y_i.

We consider three possible test statistics that we can construct based on this type of score process. Depending on their normalizing coefficients, the test statistics would follow different approximate distributions. These statistics are as follows: (i) $\max_{1<t\leqslant n} \frac{1}{\sqrt{n}}|S_t|$; (ii) $\max_{1<t\leqslant n} \left\{\max_{1<j<t} \frac{1}{\sqrt{n}}(S_t - S_j)\right\}$; and (iii) $\max_{1<t\leqslant n} \frac{1}{\sqrt{t}}|S_t|$. We can approximate nicely the first two statistics by $\sup_{0<t<1} |W(t)|$, where $W(t)$ is a standard Brownian motion process. Therefore, for a given significance level (a pre-specified

probability of false alarm), one can easily obtain the monitoring threshold, h_2, for the first two statistics by using the well-known probability distribution of such a functional of Brownian motion (see [19]),

$$\begin{aligned}\alpha = P\left(\max_{1<t\leqslant n}\frac{1}{\sqrt{n}}|S_t| > h_2\right) &\cong P\left(\max_{1<t\leqslant n}\left\{\max_{1<j<t}\frac{1}{\sqrt{n}}(S_t - S_j)\right\} > h_2\right)\\ &\cong P\left(\sup_{0<t<1}|W(t)| > h_2\right)\\ &= 1 - \frac{4}{\pi}\sum_{l=0}^{\infty}\frac{(-1)^l}{2l+1}\exp\left(\frac{\pi^2(2l+1)^2}{8h_2^2}\right),\end{aligned} \tag{6}$$

where the threshold for monitoring depends only on the probability of false alarm, that is, $h_2 = h_2(\alpha)$. A one-sided version of the first test statistic is also possible by removing the absolute value sign. Such process would have simpler tail probabilities that we can calculate from the standard normal distribution,

$$\alpha = P\left(\max_{1<t\leqslant n}\frac{1}{\sqrt{n}}S_t > h_3\right) \cong P\left(\sup_{0<t<1}W(t) > h_3\right) = 2[1-\Phi(h_3)], \tag{7}$$

where $\Phi(.)$ is the cumulative distribution function of the standard normal. Thus, we can approximate the threshold h_3 by $h_3 \cong \Phi^{-1}(1-\alpha/2)$. Darling and Erdös [20] have calculated the asymptotic distribution of the third test statistic. They showed that

$$\lim_{n\to\infty} P\left\{a(n)\max_{1\leqslant t\leqslant n} t^{-1/2}|S_t| - b(n) \leqslant h_1\right\} = \exp(-2e^{h_1}),\ -\infty < h_1 < \infty, \tag{8}$$

where $a(n) = (2\log(\log n))^{-1/2}$ and $b(n) = 2\log(\log n) + 1/2\log(\log(\log n)) - 1/2\log\pi$. From this, we can see that $\max_{1\leqslant t\leqslant n} t^{-1/2}|S_t|$ converges to infinity at the very slow rate of $(\log(\log n))^{1/2}$ as $n\to\infty$. The practical implication of this is that increasing the monitoring time horizon (truncation point) does not have a large effect on the threshold. Furthermore, the direct use of Equation (8) for the threshold gives very conservative tests, as the convergence of the maximum value is so slow. We can obtain a better approximate tail probability expression for this test statistic by approximating it through a diffusion process and then by using a formula from Vostrikova [21] for the probability distribution of the maximal functional of diffusion processes. That is, we can write

$$\alpha = P\left(\sup_{1<t\leqslant n}\frac{1}{\sqrt{t}}|S_t| > h_1\right) \cong \frac{\exp(-h_1^2/2)h_1}{\sqrt{2\pi}}\left\{\ln(n)(1-\frac{1}{h_1^2}) + \frac{4}{h_1^2} + O(\frac{1}{h_1^4})\right\}, \tag{9}$$

where $h_1 = h_1(n,\alpha)$ is a threshold that depends on both the probability of false alarm as well as on the truncation point, n.

These score tests have also been used in detecting changes in time series as well as in the context of sequential testing [22–26]. The score statistic, as opposed to the likelihood ratio statistic used in RA-CUSUM, maintains the familiar interpretation of expected minus observed, and in addition it does not require specification of an alternative hypothesis. We can now define the following four tests (monitoring procedures) based on the above three score test statistic processes.

1. Test 1: Signal an alarm if for some $1 < t \leqslant n$

$$\mathrm{STAT}_1(t) = \frac{1}{\sqrt{t}}|S_t| \geqslant h_1(n,\alpha).$$

   If $\mathrm{STAT}_1(t) < h_1(n,\alpha)$ for all $t \leqslant n$, then no evidence against the null hypothesis has been found, and the monitoring process has to be restarted.
2. Test 2: Signal an alarm if for some $1 < t \leqslant n$,

$$\mathrm{STAT}_2(t) = \frac{1}{\sqrt{n}}|S_t| \geqslant h_2(\alpha).$$

   If $\mathrm{STAT}_2(t) < h_2(\alpha)$ for all $t \leqslant n$, then no evidence against the null hypothesis has been found, and the monitoring process has to be restarted.

3. Test 3: A one-sided version of Test 2 can be obtained by removing the absolute value from the STAT_2, that is, by monitoring

$$\text{STAT}_3(t) = \frac{1}{\sqrt{n}} S_t$$

with threshold $h_3(\alpha)$.

4. Test 4: Signal an alarm if for some $1 < t \leqslant n$

$$\text{STAT}_4(t) = \max_{1<j<t} n^{-1/2}\left(S_t - S_j\right) \geqslant h_2(\alpha),$$

otherwise conclude that there is no evidence of a change in performance, and the monitoring process must restart.

Although Equation (6) for calculating the threshold h_2 for a pre-specified false alarm rate, α, is in the form of an infinite series, often the first five or so terms of the summation are enough to give an accurate result. For convenience of the users, we report in Table I some values of h_2 for several commonly used values of false alarm probability, α. For Test 1, by using Vostrikova's formula (9), we can see, for example, that the truncation point $n = 9600$, which we will use in the simulation section later on, gives for $\alpha = 0.1, 0.05$, and 0.01 the approximate thresholds $3.05, 3.30$, and 3.79, respectively. Doubling the monitoring horizon to $n = 19,200$ would give approximate thresholds $3.08, 3.32$, and 3.81, respectively. It is also worth pointing out that, in order to avoid erroneous early stopping, it may be advisable to start monitoring after at least $n_0 = 10$ observations have been accrued. This would help the procedure perform better specially in the case of Test 1, in which the maximum may be attained early in the series being monitored.

Next, we describe the risk-adjusted CUSUM of [16], designed to monitor for a change with underlying hypotheses of the form (3).

5. RA-CUSUM: Signal an alarm if for $t > 1$

$$Z_t = \text{Max}\left[0, Z_{t-1} + W_t\right] > h, \tag{10}$$

where $Z_0 = 0$, and W_t is the likelihood ratio contribution of the tth patient, defined by

$$W_t = \begin{cases} \log\left[\frac{(1-\pi_t+R_0\pi_t)R_A}{(1-\pi_t+R_A\pi_t)R_0}\right] & \text{if } y_t = 1, \\ \log\left[\frac{(1-\pi_t+R_0\pi_t)}{(1-\pi_t+R_A\pi_t)}\right] & \text{if } y_t = 0. \end{cases} \tag{11}$$

The threshold value h is a function of the ARL under the null hypotheses of no change. We can compute this critical value by using an exact formula that requires numerical integration, by using approximate Markov chain techniques, or by using Monte Carlo simulations. For detecting improved performance, Steiner *et al.* [16] suggested using an updated formula for detecting decreases in the treatment failure rate, that is, $R_A < R_0$, so that the RA-CUSUM with

$$Z_t = \min\left[0, Z_{t-1} - W_t\right] \tag{12}$$

will accumulate negative values.

The RA-CUSUM procedure defined in the previous paragraph is based on Page's original CUSUM procedure [27], and as such it is asymptotically optimal in the sense of minimizing the delay of alarm, given that the time of change (the change point) has been passed [28, 29]. The test is, however, open ended, which means that there is no maximum sample size, although it will eventually stop with probability 1 at some finite sample.

Table I. Threshold values h_2 and h_3 for various probabilities of false alarm, α.

α	0.01	0.05	0.1	0.15	0.20	0.25	0.30	0.35	0.40	0.45
h_2	2.8070	2.2414	1.9600	1.7805	1.6448	1.5341	1.4395	1.3562	1.2812	1.2126
h_3	2.5758	1.9600	1.6449	1.4395	1.2816	1.1503	1.0364	0.9346	0.8416	0.7554

In contrast, the proposed sequential Tests 1–4 have a maximum sample size n (also known as monitoring horizon or truncation point). One way of choosing the truncation point n is to take the ARL of the RA-CUSUM under the null hypothesis as initial guess and fine-tune it by using Monte Carlo simulations in order to attain a desired power. An alternative approximate method for choosing n is based on calculations similar to sample size calculations in the nonsequential testing setup. To illustrate the latter, consider Test 4 and fix the level of significance to α. As the approximating large sample distribution is now the distribution of $\sup_{0<t<1}|W(t)|$, that is, the supremum of the standard Brownian motion's absolute value on the interval (0, 1), this distribution will replace the standard normal distribution in the calculations of power for normal tests. Furthermore, assume that we wish to attain a power of $1-\beta$ in detecting an alternative odds ratio of R_A with a delay of $n\delta$, $0<\delta<1$, observations after the change point $\tau=\gamma n$, $0<\gamma<1$. As the scores $y_i-\pi_i$ depend on the Parsonnet score x_i as well, where x_i is assumed to be a random variable, we have to choose a value for x_i. One can take the average value of Parsonnet scores in the historical data to calculate π_i. Then routine calculations give the minimum truncation point value

$$n \geqslant \left(\frac{F_\alpha - F_{1-\beta}}{2\delta(\pi_A-\pi_0)}\right),$$

where F_α denotes the upper α-percentile of the distribution of $\sup_{0<t<1}|W(t)|$ and where we use a factor of $1/2$ as the maximum value of $\sqrt{\pi(1-\pi)}$. As an example, we take $R_0=1$, $R_A=2$, and the average historical Parsonnet score $x=40$, and at the level of $\alpha=0.05$ we wish to have power $1-\beta=0.8$ to detect this shift within $0.1n$ observations. This gives us $\delta=0.1$, $F_{0.05}=2.24$, $F_{0.8}=0.82$, and $\pi_A-\pi_0=0.169$. We get $n\geqslant 1765$ is a sufficiently large value of the truncation point. The actual power achieved can be greater or smaller depending on whether $\delta+\gamma$ is smaller or greater than 1, respectively.

3. Numerical studies

3.1. Simulation studies

We considered the model $\text{logit}(p_t)=-3.68+0.077x_t$, where the coefficients are based on the surgeon monitoring data set between 1992 and 1994 as in [16]. In the simulation study, we randomly (with replacement) sampled Parsonnet scores from the actual scores of the surgeon monitoring data (1992–1998). For the RA-CUSUM, we used $h=4.5$, and for all tests, the odds ratios were varied over $R_0=1$ and $R_A=1.5, 2$. The upper quartile Q_3 of the RA-CUSUM's run lengths was found to be 9751 under the null hypothesis of no change. Accordingly, we set the truncation point for the rest of the tests near that value, $n=9600$, and we computed $h_1(n,\alpha)=h_1(9600,0.05)=3.28$, whereas $h_2(0.05)=2.24, h_3(0.05)=1.96$, independent of n. The change point was set at the start of the monitoring process, $\tau=1$ and at time points $\tau=3000, 6000, 9000$.

Also, based on the simulation results, we plotted in Figures 1 and 2 the probability of a false alarm under a doubling odds ratio and the conditional power of all tests as functions of the change point τ, respectively. We obtained the probability of false alarms as the proportion of signals occurring before the change point τ, whereas we defined the conditional power as the proportion of signals occurring after τ but before the truncation point 9600 for the case of RA-CUSUM.

We did the Monte Carlo simulations in R (R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria) [30], and each scenario was repeated 5000 times to compute ARLs and error probabilities. In Table I, we reported the first and third quartiles (Q_1, Q_3), the median Q_2, the mean (ARL), and the maximum of the distribution of the run lengths of the RA-CUSUM. We also reported in the same table the ARLs for runs stopping on or after the change (ARL $\geqslant\tau$) and those stopping before the change (ARL $<\tau$) as well as the empirical probability of stopping before the change, computed as the fraction of stops occurring before τ out of the 5000 runs. For Tests 1–4, we reported the same results as for RA-CUSUM, together with the probabilities of type I errors and powers, in Tables II–IV.

From Table I, we see that for the RA-CUSUM, the probability of stopping before the change point can be as high as 70% in certain configurations. Also, the unconditional ARLs are well below the change point itself when the change is not at the beginning of the series being monitored. This is clear from the column headed by ARL in Table I, where ARL = 2581, 4012, 5332 for $\tau=3000, 6000, 9000$, respectively. These ARLs are before their respective change points. However, the conditional ARLs

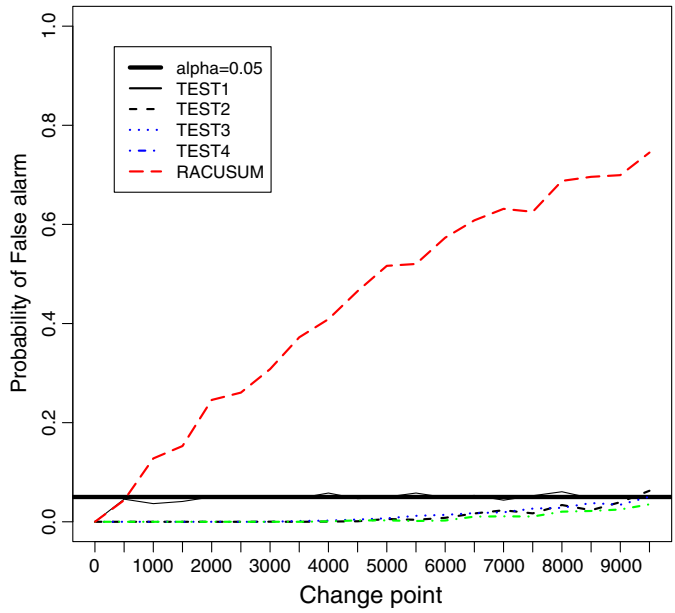

Figure 1. Probability of false alarm of all tests as functions of the time of change τ for doubling odds ratio and truncation point of $n = 9600$ for Tests 1–4.

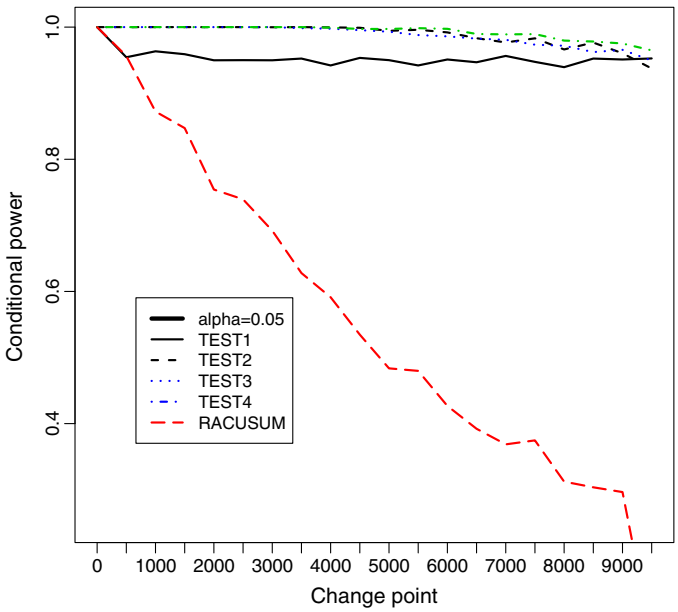

Figure 2. Conditional power of all tests as functions of the time of change τ for doubling odds ratio and truncation point of $n = 9600$ for Tests 1–4.

of the RA-CUSUM, given that the change point has been passed, are the shortest among all tests reported here (see the columns headed by ARL $\geqslant \tau$). For instance, the conditional run lengths are ARL $= 206, 3183, 6188, 9184$ for changes of magnitude $R = 2$ occurring at $\tau = 1, 3000, 6000, 9000$, respectively. This is consistent with the optimality properties pointed out by [27] and [29].

Tables II–VI show that Tests 1–4 indeed have good control over their type I errors (probabilities of false alarm). Test 1's ARLs are the shortest, but its power is lower than the power of the other truncated tests. Tests 2–4 are comparable in terms of their ARLs; although Test 4 is quite conservative, it is best in detecting late changes. It is obvious that these new methods detect change with greater delay than RA-CUSUM, but their power is high, and we can trust that they stop for genuine changes most of the time.

Table II. Simulated quartiles Q_1, Q_2, and Q_3, maximum and average of the run lengths of RA-CUSUM with $h = 4.5$ for detecting changes in odds ratio R at times $\tau = 1, 3000, 6000, 9000$. ARL $\geqslant \tau$, ARL $< \tau$, and P_τ are the average run lengths (ARLs) for runs stopping on or after τ and those stopping before τ, and the probability of stopping before τ, respectively.

R	Q_1	Q_2	ARL	Q_3	Max	ARL $\geqslant \tau$	ARL $< \tau$	P_τ
				$\tau = 1$				
1.0	2063	4824	6967	9751	53280	6967	NA	0.00
1.5	236	421	546	725	3420	546	NA	0.00
2.0	114	174	206	263	1258	206	NA	0.00
				$\tau = 3000$				
1.5	2055	3204	2798	3513	6541	3521	1396	0.34
2.0	2079	3089	2581	3188	3928	3183	1457	0.35
				$\tau = 6000$				
1.5	2086	4952	4322	6339	9387	6523	2651	0.57
2.0	1737	4659	4012	6129	6721	6188	2547	0.60
				$\tau = 9000$				
1.5	1908	4826	5209	9140	13140	9528	3462	0.71
2.0	2227	5181	5332	9077	9781	9184	3666	0.70

Table III. Simulated type I error, quartiles Q_1, Q_2, and Q_3, maximum and average of the run lengths of Test 1 with truncation point $n = 9600$ and $\alpha = 0.05$ for detecting real changes in odds ratio R at times $\tau = 1, 3000, 6000, 9000$. ARL $\geqslant \tau$, ARL $< \tau$, and P_τ are the average run lengths (ARLs) for runs stopping on or after τ and those stopping before τ, and probability of stopping before τ, respectively.

R	$\hat{\alpha}$	Q_1	Q_2	ARL	Q_3	Max	ARL $\geqslant \tau$	ARL $< \tau$	P_τ
					$\tau = 1$				
1.0	0.055	9600	9600	9114	9600	9600	9114	NA	0.000
1.5	1.000	324	678	811	1181	3702	811	NA	0.000
2.0	1.000	103	203	248	344	1164	248	NA	0.000
					$\tau = 3000$				
1.5	1.000	4494	5070	5016	5747	8789	5235	218	0.044
2.0	1.000	3727	3996	3856	4242	5849	4038	191	0.047
					$\tau = 6000$				
1.5	0.758	7974	8791	8317	9570	9600	8721	432	0.049
2.0	1.000	6964	7324	6979	7668	8930	7382	648	0.060
					$\tau = 9000$				
1.5	0.058	9600	9600	9120	9600	9600	9599	590	0.053
2.0	0.060	9600	9600	9298	9600	9600	9596	1273	0.036

Table IV. Simulated type I error, quartiles Q_1, Q_2, and Q_3, maximum and average of the run lengths of Test 2 with truncation point $n = 9600$ and $\alpha = 0.05$ for detecting real changes in odds ratio R at times $\tau = 1, 3000, 6000, 9000$. ARL $\geqslant \tau$, ARL $< \tau$, and P_τ are average run lengths (ARLs) for runs stopping on or after τ and those stopping before τ, and probability of stopping before τ, respectively.

R	$\hat{\alpha}$	Q_1	Q_2	ARL	Q_3	Max	ARL $\geqslant \tau$	ARL $< \tau$	P_τ
					$\tau = 1$				
1.0	0.050	9600	9600	9487	9600	9600	9487	NA	0.000
1.5	1.000	1720	2035	2080	2409	4433	2080	NA	0.000
2.0	1.000	926	1052	1076	1202	1853	1076	NA	0.000
					$\tau = 3000$				
1.5	1.000	4565	5025	5094	5552	8553	5094	NA	0.000
2.0	1.000	3850	4068	4081	4302	5346	4081	NA	0.000
					$\tau = 6000$				
1.5	0.945	7473	8070	8076	8684	9600	8088	5238	0.004
2.0	1.000	6768	7037	7056	7355	8660	7073	5149	0.009
					$\tau = 9000$				
1.5	0.077	9600	9600	9487	9600	9600	9591	6948	0.039
2.0	0.184	9600	9600	9484	9600	9600	9568	7269	0.037

Table V. Simulated type I error, quartiles Q_1, Q_2, and Q_3, maximum and average of the run lengths of Test 3 with truncation point $n = 9600$ and $\alpha = 0.05$ for detecting real changes in odds ratio R at times $\tau = 1, 3000, 6000, 9000$. ARL $\geqslant \tau$, ARL $< \tau$, and P_τ are average run lengths (ARLs) for runs stopping on or after τ and those stopping before τ, and probability of stopping before τ, respectively.

R	$\hat{\alpha}$	Q_1	Q_2	ARL	Q_3	Max	ARL $\geqslant \tau$	ARL $< \tau$	P_τ
					$\tau = 1$				
1.0	0.060	9600	9600	9451	9600	9600	9451	NA	0.000
1.5	1.000	1467	1760	1807	2088	4034	1807	NA	0.000
2.0	1.000	809	931	953	1084	1828	953	NA	0.000
					$\tau = 3000$				
1.5	1.000	4318	4787	4824	5262	7510	4827	2791	0.002
2.0	1.000	3711	3933	3952	4179	5239	3953	2821	0.001
					$\tau = 6000$				
1.5	0.966	7250	7793	7805	8392	9600	7871	4659	0.021
2.0	1.000	6652	6935	6933	7220	8555	6966	4734	0.015
					$\tau = 9000$				
1.5	0.120	9600	9600	9447	9600	9600	9583	6803	0.049
2.0	0.257	9581	9600	9380	9600	9600	9552	6128	0.050

Table VI. Simulated type I error, quartiles Q_1, Q_2, and Q_3, maximum and average of the run lengths of Test 4 with truncation point $n = 9600$ and $\alpha = 0.05$ for detecting real changes in odds ratio R at times $\tau = 1, 3000, 6000, 9000$. ARL $\geqslant \tau$, ARL $< \tau$, and P_τ are average run lengths (ARLs) for runs stopping on or after τ and those stopping before τ, and probability of stopping before τ, respectively.

R	$\hat{\alpha}$	Q_1	Q_2	ARL	Q_3	Max	ARL $\geqslant \tau$	ARL $< \tau$	P_τ
					$\tau = 1$				
1.0	0.028	9600	9600	9555	9600	9600	9555	NA	0.000
1.5	1.000	1820	2112	2176	2483	4615	2176	NA	0.000
2.0	1.000	988	1122	1144	1283	2152	1144	NA	0.000
					$\tau = 3000$				
1.5	1.000	4447	4780	4821	5163	7303	4821	NA	0.000
2.0	1.000	3776	3936	3946	4113	4899	3946	NA	0.000
					$\tau = 6000$				
1.5	1.000	7251	7648	7638	8067	9600	7651	4934	0.005
2.0	1.000	6630	6848	6830	7059	7791	6847	5395	0.011
					$\tau = 9000$				
1.5	0.120	9600	9600	9499	9600	9600	9589	6920	0.034
2.0	0.311	9563	9600	9476	9600	9600	9549	6896	0.027

From Figures 1 and 2, we can see that the probability of a false alarm with the RA-CUSUM increases, whereas its conditional power decreases as the time of change, τ, increases.

3.2. *Application to real data*

We illustrate the methodologies by using data collected at a UK center for cardiac surgery. The data consist of patients' pre-operative covariate information such as age, gender, history of hypertension, which were summarized as patient Parsonnet scores, as well as surgery date and identification numbers for the surgeons. In this application, the outcome of interest is the 30-day post-operative mortality rate. The data relating to the period between 1992 and 1994 were used to build a baseline logistic regression model of the form $\log\left(\frac{p_t}{1-p_t}\right) = -3.68 + 0.077x_t$. We then monitored mortality rates of patients operated by seven of the surgeons during the period 1994–1998. The monitoring process using RA-CUSUM was reported graphically in [16]; therefore, here we only report the monitoring processes of the new procedures, Tests 1–4 (see Figures 3–6). The monitoring horizon was taken to be $n = 9600$, which is the empirical in-control ARL of the RA-CUSUM, and the significance level was set to $\alpha = 0.05$, thus obtaining $h_1 = h_1(n, \alpha) = h(9600, 0.05) = 3.26$ from formula (9) and $h_2 = 2.24, h_3 = \pm 1.96$ from Table I. It seems that Tests 2–4 are consistent with the results of [16], whereas Test 1 signals a change for surgeons 1 and 2.

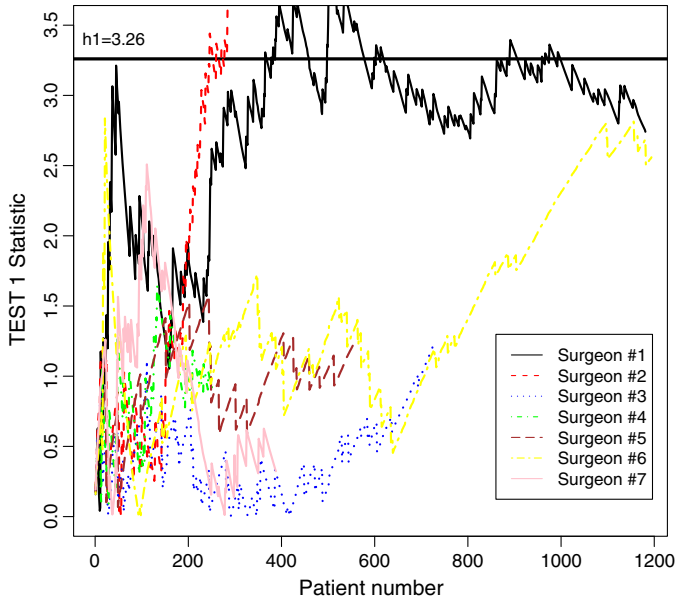

Figure 3. Monitoring of the seven surgeons by using Test 1, $h_1(9600, 0.05) = 3.26$, and data between 1994 and 1998.

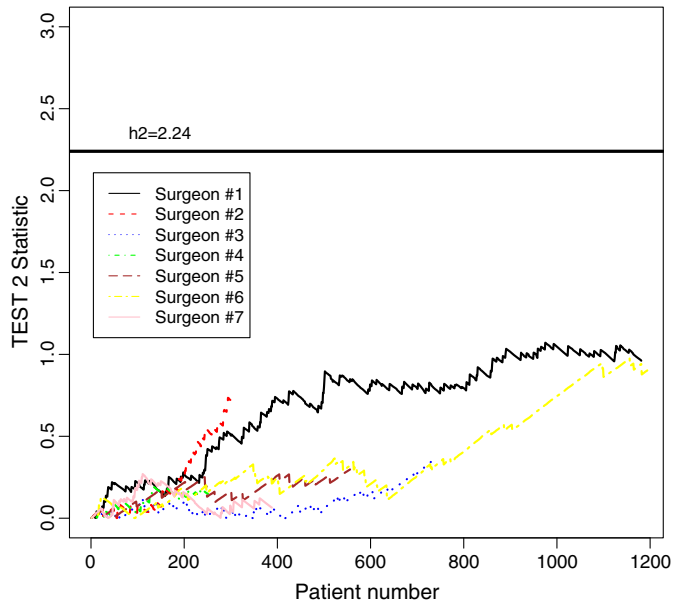

Figure 4. Monitoring of the seven surgeons by using Test 2, $h_2(0.05) = 2.24$, and data between 1994 and 1998.

4. Discussion

The various tests presented in this paper have different early stopping and error probability characteristics. RA-CUSUM stops the fastest if the change point has been passed, but because of frequent erroneous stops before the change, the user cannot tell if the signal indicates a real change or not. The new tests, Tests 1–4, have roughly fixed type I error rates as in the classical statistical testing theory. This makes the alarms more reliable signals of genuine change. The price for this is an increased delay in detection. Hence, in these sequential methods, the two important properties of fast stopping and error rate have to be considered when choosing a monitoring process for a given application. Among the new tests proposed here, Test 1 stops the fastest for large changes, but Tests 2 and 3 have more power when the magnitude of the change is small. Test 4 is best if late changes in the monitoring horizon are of concern.

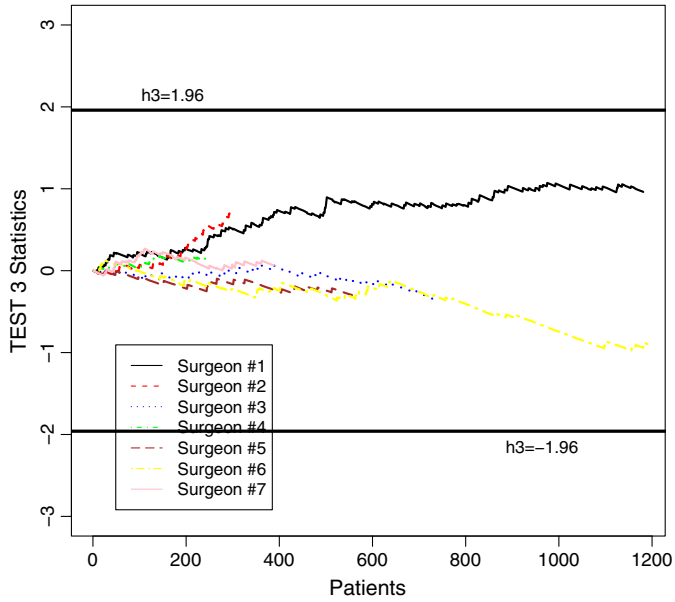

Figure 5. Monitoring of the seven surgeons by using Test 3, $h_3(0.05) = \pm 1.96$, and data between 1994 and 1998.

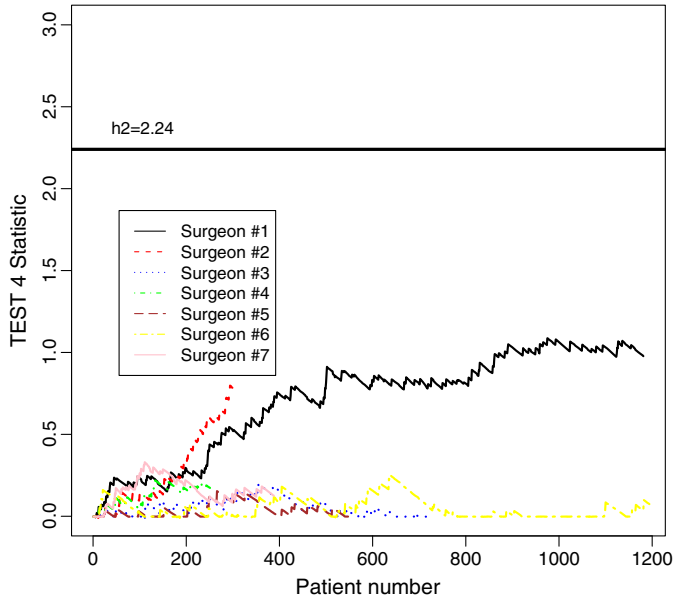

Figure 6. Monitoring of the seven surgeons by using Test 4, $h_2(0.05) = 2.24$, and data between 1994 and 1998.

Acknowledgements

The authors would like to thank the editor and the anonymous reviewers for their constructive comments, which led to the improvement of the original manuscript. This research was supported by the Natural Science and Engineering Research Council of Canada (NSERC).

References

1. Lie R, Heuch I, Irgens L. A new sequential procedure for surveillance of Downs-Syndrome. *Statistics in Medicine* 1993; **12**(1):13–25.
2. Inquiry. BRI Inquiry Panel. Learning from Bristol: the report of the public inquiry into children's heart surgery at the Royal Infirmary 1984–1995, London, UK: The stationery Office, 2001. Available from http://www.bristol-inquiry.org.uk/*final_report*/.

3. Inquiry. Shipman Inquiry. *The First Report*, London, UK: The stationery Office, 2002. Available from http://www.the-shipman-inquiry.org.uk/reprt/asp.
4. Spiegelhalter D, Grigg O, Kinsman R, Treasure T. Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. *International Journal for Quality in Health Care* 2003; **15**(1):7–13.
5. Lovegrove J, Valencia O, Treasure T, Sherlaw-Johnson C, Gallivan S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* 1997; **350**:1128–1130.
6. Lovegrove J, Sherlaw-Johnson C, Valencia O, Treasure T, Gallivan S. Monitoring the performance of cardiac surgeons. *Journal of the Operational Research Society* 1999; **50**(7):684–689.
7. Poloniecki J, Valencia O, Littlejohns P. Cumulative risk adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *British Medical Journal* 1998; **316**(7146):1697–1700.
8. Cook D, Steiner S, Cook R, Farewell V, Morton A. Monitoring the evolutionary process of quality risk-adjusted charting to track outcomes in intensive care. *Critical Care Medicine* 2003; **31**(6):1676–1682.
9. Grigg O, Farewell V. A risk-adjusted Sets method for monitoring adverse medical outcomes. *Statistics in Medicine* 2004; **23**(10):1593–1602.
10. Woodall W. The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology* 2006; **38**(2):89–104.
11. Grigg OA, Spiegelhalter DJ. An empirical approximation to the null unbounded steady-state distribution of the cumulative sum statistic. *Technometrics* 2008; **50**(4):501–511.
12. Grigg OA, Spiegelhalter DJ. Clinical surveillance and patient safety. In *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects,* Cambridge Books Online. Cambridge University Press, 2010.
13. Lai TL. Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society Series B* 1995; **57**:613–658.
14. Mei Y. Is average run length to false alarm always an informative criterion? *Sequential Analysis* 2008; **27**(4):354–376.
15. Gombay E, Serban D. Monitoring parameter change in AR(p) time series models. *Journal of Multivariate Analysis* 2009; **100**(4):715–725.
16. Steiner SH, Cook RJ, Farewell VT, Treasure T. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* 2000; **1**(4):441–452.
17. Parsonnet V, Dean D, Bernstein A. A method Of uniform stratification of risk for evaluating the results of surgery in acquired adult heart-disease. *Circulation* 1989; **79**(6):3–12.
18. Wald A. *Sequential Analysis*. John Wiley and Sons: New York, 1947.
19. Borodin AN, Salminen P. *Handbook of Brownian Motion—Facts and Formulae*. Birkhauser: Basel, 1996.
20. Darling DA, Erdös P. A limit theorem for the maximum of normalized sums of independent random variables. *Duke Mathematical Journal* 1956; **23**:143–155.
21. Vostrikova LJ. Detection of a 'disorder' in a Wiener process. *Theory of Probability and its Applications* 1981; **26**:356–362.
22. Gombay E. Sequential testing of composite hypothesis. In *Limit Theorems in Probability and Statistics II*, Berkes I, Csáki E, Csörgő M (eds), 2002; 107–125.
23. Gombay E. Parametric sequential tests in the presence of nuisance parameters. *Theory of Stochastic Processes, Kiev* 2002; **8**(24):107–118.
24. Gombay E, Serban D. Monitoring parameter change in AR(p) time series models. In *Statistics Centre Technical Reports 05.04*. The University of Alberta, Edmonton: Canada, 2005.
25. Gombay E, Hussein A. A class of sequential tests for two-sample composite hypotheses. *Canadian Journal of Statistics* 2006; **34**(2):217–232.
26. Gombay E. Sequential change-point detection and estimation. *Sequential Analysis* 2003; **22**(3):203–222.
27. Page ES. Continuous inspection schemes. *Biometrika* 1954; **41**:100–115.
28. Lorden G. Procedures for reacting to a change in distribution. *Annals of Mathematical Statistics* 1971; **42**:1897–1908.
29. Moustakides GV. Optimal stopping times for detecting changes in distributions. *Annals of Statistics* 1986; **14**(4): 1379–1387.
30. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria 2011. http://www.R-project.org, ISBN 3-900051-07-0.